A COMPUTATIONAL LINGUISTICS APPROACH TO CONCEPTUAL INFORMATION PROCESSING *

Dr. Inger Bierschenk

Research associate. Dept. of Educational & Psychological Research,
University of Lund-Malmö, Sweden

Introduction

The purpose of this paper is to present a linguistic model which can be used for the processing of scientific information. The model has developed through several years of research and experiments in information mediation, structuring, organisation and dissemination in such social and humanistic sciences that have been of interest to Swedish educational research departments.

It has been apparent that a so-called "information overflow" and "information explosion" are irrelevant concepts to designate the state of affairs. There exists an ever growing data flow but a considerable lack of information, for example among many professional categories requiring information, such as university teachers and researchers. This so-called "information frustration" is likely to depend on the meaning an individual abstracts or infers from the data, e.g., in the interpretation of concepts and conceptual relations mediated in document titles. In computerized IoD search strategies there is a more serious barrier involved. If the logic of the computer or the system is unknown, there is a high risk of uncertainty as to the real existence of the documents, which makes for an abstractness of an even higher order. Moreover, when the information mediator conceptualizes in a way different from the searcher, which is often the case in social science research, it will prove difficult to arrive from data to information.

There are many science fields dealing with information, e.g., library science, information science, artificial intelligence, computational linguistics and the emerging field of cognitive science. All of them are concerned with the study of mechanisms that make possible representation of information derived from symbols. Since computational linguistics focuses on the study of language use aided by computers, there exists a link to

* IATUL 9th biannual conference, Lausanne, 1981

library and information science. For two or three decades great efforts have been made in the field of linguistics, especially its computational variety, to meet the needs in information science of structuring and deriving meaningful results in the processing of information. One basis goal for IoD systems should be the dynamic structuring, since information is characterized by the structuring and re-structuring of data, and thus subject to constant change. But there are strong traditions in linguistics as well as in library science, which seem to survive, despite the new technology of the electronic revolution. In fact, the symbol handling machine contributes to the conservation of this tradition. In the following I shall try to explain the reason why by means of some general principles governing the structuring of information.

/c

Some principles in the structuring of information

There are different forms of representation with different goals. If the goal is, as it was once in the history of library organisation, to store "all" documented (written) information available, it seems natural that a philosophy of universal classification lies behind the structuring of the items. A well-known and common type is made up of the principle of hierarchy. Such a system has only a small potential for quick and easy adaptation to a particular structure in looking for new information. Concretely, once a book is put on a shelf it is bound to its place. A more flexible organization is provided by facet classification, through the lateral relationships made explicit.

Recent efforts to structure information have been made through different kind of networks. The starting point was Ross Quillian´s[1] memory model assuming that human cognition is associatively structured. The so-called semantic network builds on concepts, represented as node labels which have links to their attributes (properties). In information search systems these networks are used in answering questions about facts in a data base. Such a fact, e.g. a property of a certain concept, may not directly lead to an answer to facts about a concept of higher order.

Later constructions of semantic networks make clear that this representation is founded on the classical-philosophical way of structuring the world, similar to generic and other lexical-semantic relationships. It may therefore be called static. Thus the network principle is based on relationships within each concept. The structure does not account for contextual relations such as they emerge in natural language sentences, i.e. the dynamic properties of information. An attempt in this direction is the PRECIS system which tries to build in structural relationships in the coding but still seems to rely very much on traditional classification principles.

In computer science, especially among those artificial intelligence researchers dealing with data base management, there is a trend to use the network principle in a restricted frame or context by the construction of so-called topic hierarchies[2] in order to prevent the associations from "exploding" in the information search procedure. The AI researchers are mainly developing mechanisms for logical deduction, i.e. the application of rules of inference to statements made in a formal language, whose "semantics" is well specified. Thus since the intelligence of the computer is restricted, dynamic structuring is not possible. It is an automatization of the philosophy and logic built into classic information structuring, and, unfortunately, the analogy made concerning theories of the structuring and functioning of human cognition is one main reason for the maintenance of the traditional reasoning in the construction of information processing systems.

Simultaneously, another view of structuring, the schema principle, is discussed among researchers in cognitive psychology. This model emanates from Sir Fredric Bartlett´s statements on the human memory as a propositional structure[3]. As opposed to semantic networks whose organizers have to account for how many "semantic primitives" are needed for a synthetic formation of concepts, a structure based on schemata tries to explore the advantages of an adaptively operating process. The schema utilizes the relations between abstractions. This means that information need not be explicit as in the network, but implicit, i.e. embedded in the structure.

Formats of representation

From a linguistic point of view a model based on propositions would be represented as a Noun_1 -v- Noun_2 model, or in its more logical variety, as a predicate-argument statement: $v \left[N_1, \ldots \right]$. This representation format is often used in the processing and computing of natural language data. The inferences are then drawn to the semantic or the logical structure of the text. The text is in these cases one sentence. However, when the purpose is to inform about a sequence of events, as e.g. a scientific process, a process oriented model is required. For Indo-European languages, the Agent-action-Object paradigm seems to be general. This model is dynamic and takes direction and intentionality into account and can be used in psychologic studies of information processed from whole texts[4]. By means of the

AaO model latent dimensions in the information structure can be detected[5]. Inferences are made to the relational structure between concepts and conceptualizations.

Theoretical starting-points for an information processing experiment

Starting with the assumption that the document title is the first and often only contact an information searcher has with a scientific text in the process of judging its content, its language structure would be conceived as being the communicative surface between author and reader. This intermediate language is in the experiment assumed to represent an abstraction of all "scientific events" reported in the text as a whole.

If scientific reporting is concerned with establishing causal connections between events[6], these should be detected by means of a model whose components are assigned variables representing scientific entities instead of linguistic or psychological ones. Such a model has been developed by the name of the Problem-method-Goal paradigm assuming these concepts being central to research work[7]. The PmG model thus represents an abstract proposition and as such its characteristics differ in operationalization from more natural-language adapted models. This may be illustrated with an example from a transformational stage. From an interview about research there could be a statement like

I have analyzed titles for several months (1)

where the researcher is present (I) and tells about a process that has happened (verb forms) and also for how long time. When the same author writes a report about his research the time specification will not be present in the title (there is no "here and now") and his person is implicit:

An analysis of titles (2)

A transformation has taken place. In order to mark the object of his study the preposition "of" functions as pointer. The activity has been transformed to one concept designating some kind of research or investigation method. Since research does not handle concrete object but instead problems, /ts it is appropriate to change the label "object" in favour of the label "problem". Thus m and P have been discerned. The G component is marked through the preposition "for". There is also an optional component in the model when a research instrument is involved. The preposition to mark instruments is "with". This is described in more detail in I. Bierschenk[8].

The operationalization by means of an algorithm that automatically codes titles in accordance with this conceptual information structure builds on a few but general principles. I believe that these principles can be used to explain the processing of natural language and the procedure of mediating information between man and the machine. If the interpretation of language data is based on cognitive functions such as they are studied by cognition oriented scientists, much of the confusion between the outcome based on formal logic and the logic of conception may disappear.

Among others, Piaget and Inhelder[9] have made extensive experiments showing that human cognition seems to be spatially organized. The child´s conception develops from a simple demarcation of objects to an understanding of two-dimensional relations of several kinds. Adults learn to relate multidimensional phenomena and seem to use these organizing principles in a kind of coordinate system for the orientation in space and time. The basic hypothesis underlying the algorithm is that there are cues in the intermediate structure of the scientific title that relate the concepts and organize them in such a way that a cognitive structure can be detected (i.e. re-cognized). The prepositions are therefore used here as functors, structuring the scientific conceptualization by means of an unambiguous organization. Based on the intentionality of the PmG model and the visually signalled explicitness of conceptual demarcation the following computational linguistic model has governed the algorithmic analysis.

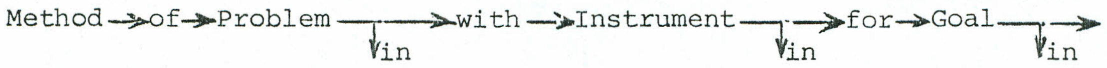

Figure 1. General principles of an algorithm for conceptual recognition

Some prepositions point to concepts, indicating the direction (intention) of the research activity (here called "intentional" prepositions). Others procedurally demarcate, i.e. give an explicit specification of the concept in question. This algorithmic definition of the relationship has created the name "extensional" prepositions. A more thorough discussion is found in I. Bierschenk[8] , where also the rules for decoding at several levels are listed and described.

Patterns in an experimental data base

For the testing of the algorithm a representative sample of 9,000 bibliographic descriptions has been used. The sample is based on a definition of "researcher". Works produced of these researchers constitute the scientific title collection.

According to the analytical model employed, the conceptualizations (as represented by positions in the schema) may be extended in varying degrees. Experimental results show that the most characteristic feature in this science field is the Problem component, single or with one extension. Because of the intentionality inherent in the model, this pattern should be interpreted as expressing an implicit intentionality. If two components appear it is most likely a Problem and a Method. The presence of the method should be interpreted as an explicit expression of intentionality, as in research itself. The more complex the patterns are the less often they appear. There also seems to be a tendency that many extensions prevent "transitions" between components.

A validation of this conceptual structure has been performed by correlating the pattern types with types of document. For example, textbooks and other kinds of monographs are less complex than formal articles, which in turn are less complex than research reports. Thus three general structures appear in the material, namely (1) explicit intentionality, (2) implicit intentionality + second degree extensionality, and (3) explicit intentionality + first degree extensionality.

Conceptual information in scientific titles

Apart from some titles (about .005%) with a low abstraction level compared with expectancies and structural logic built into the algorithm, the conceptual decoding has resulted in some data registers (files) corresponding to the various components of the model. These registers are now funcionally related to each other because of the non-philosophical classification. The schema model as a structuring principle also reveals such dimensions that a manual analysis could have performed only with difficulty. To illustrate the difference I would like to discuss the title

```
   Integration of children with handicaps                        (3)
        m              P            I
```

According to the model "children" is the problem dealt with in this study and the methods and techniques are abstracted in the concept "integration": the different steps to take are not explicit. A linguistic analysis, when the interpretation model is anchored in natural language variation, would regard the with-phrase, i.e. the concept "handicaps" as associated to "children", which classifies it as a property. The research-oriented information model, however, assumes this property to be instrumental. In reality, this should be the reason why the author expresses or explicates the instrument. It is likely that the integration methodology is different for those children. Therefore, it was of special interest in this report.

The schematic generality, because of the fixed positions of the components detects the variability, i.e. the variations among the values assumed under each variable. Thus "integration" can be a practical way of handling the children and also a method of study. Examples of variability of methods generated are "research" (incorporating several actions), "reflections" (a way of reporting one´s result), "handbook" (a kind of methodological strategy in educating research students). Consider also

$$\underset{m}{\text{Goals}}\ \text{for}\ \underset{G}{\text{teacher training}} \qquad \underset{P}{\text{Goals}}\ \text{in teacher training} \qquad (4)$$

The concept "goals" is a method in the first case, representing activities among these researchers involving goal description. It has a clear methodological meaning within educational technology in which teacher training is the overall goal. The preposition "for" recognizes the first "goal" as a method, whereas the preposition "in" codes it as a problem, since an explicit intentionality does not exist in the second case. Thus goals in teacher training just specifies the context within which a certain problem is dealt with. That "goals" is a noun is not of any import in the funcionally oriented registers. The author may discuss the same goals in the two titles, but from different viewpoints, from different functional domains. This also makes the following title functionally communicative to the information searcher

School for the 80´s (5)

It was written 20 years ago when the 80´s still belonged to the future. The Method component is here given a broader meaning, since the school may also be seen as an instrument. Method and instrument are components which can form method-(instrument)-goal hierarchies in relation to the degree of complexity in the desired goals. In order to reduce method and instrument to a simple concept the term "means" is used. In the light of the theoretical assumption and knowledge of this authors activities and field of inquiry in Swedish educational research, I believe that the proposed interpretation can be validated.

References

1. QUILLIAN, R. Semantic memory. In: Minsky, M. ed. Semantic information processing. Cambridge, MIT Press, 1968, pp. 216-270.

2. SCHUBERT, L.K., GOEBEL, R.G. & CERCONE, N.J. The structure and organization of a semantic net for comprehension and inference. In: Findler, N.V. ed. Associative networks. New York, Academic Press, 1979, pp. 121-175.

3. BARTLETT, F.C. Remembering. A study in experimental and social psychology. London, Cambridge University Press, 1932.

4. BIERSCHENK, B. & BIERSCHENK, I. A system for a computer-based content analysis of interview data. Lund, Gleerup, 1976.

5. BIERSCHENK, B. A new approach to psychometric problems in the analysis of pre-numeric data. Didakometry, 55, 1977.

6. SAGER, N. Perspective paper: Computational linguistics. In: Walker, D., Karlgren, H. & Kay, M. eds. Natural language in information science. Stockholm, Skriptor, 1977.

7. BIERSCHENK, B. Perception, strukturering och precisering av pedagogiska och psykologiska forskningsproblem på pedagogiska institutioner i Sverige. /Perception, structuring and definition of educational and psychological research problems on departments of education research in Sweden./ Pedagogisk-psykologiska problem, 254, 1974. /In Swedish/

8. BIERSCHENK, I. Intermediate language structure. A method for the generation of a language for representing scientific information. Ph.D. Dissertation. Göteborg, Department of Computational Linguistics, 1980.

9. PIAGET, J. & INHELDER, B. The child´s conception of space. London, Routledge & Kegan Paul, 1956.